DOCUMENT RESUME

ED 351 309 SP 034 142

AUTHOR	Boothroyd, Roger A.; And Others
TITLE	What Do Teachers Know about Measurement and How Did They Find Out?
PUB DATE	Apr 92
NOTE	24p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 1992).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Competence; Grade 7; Grade 8; Grading; Higher Education; Junior High Schools; *Knowledge Level; Mathematics Teachers; Mathematics Tests; *Measurement Techniques; Required Courses; Science Teachers; Science Tests; Student Evaluation; Teacher Certification; Teacher Education; *Teacher Made Tests; *Test Construction; *Test Theory
IDENTIFIERS	*Teacher Knowledge

ABSTRACT

Given the frequency with which teachers use self-developed tests to evaluate students, and given the paucity of requirements related to developing measurement competencies, some educators and measurement specialists question the adequacy of teachers' training in and knowledge of measurement principles. This study assesses teachers' measurement training and the extent to which their measurement knowledge is adequate to develop quality classroom tests. Forty-one 7th- and 8th-grade science and mathematics teachers were assessed using a 65-item multiple-choice test and an interview protocol. Participants were asked to identify violations of item writing principles in 32 multiple-choice and completion items. Three questions were addressed: (1) What was the nature and extent of measurement training? (2) What measurement knowledge and skills did these teachers possess? and (3) What teacher characteristics are related to their measurement knowledge? Results indicated that teachers' knowledge of measurement was insufficient, probably at least partially due to inadequate training; and that teachers frequently tested students with their own tests and placed more weight on students' scores on these tests when assigning end-of-course grades than on other forms of assessment. (LL)

What Do Teachers Know About Measurement and How Did They Find Out?

Roger A. Boothroyd

Robert F. McMorris

Robert M. Pruzek

State University of New York at Albany

A paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA, April, 1992.

What Do Teachers Know About Measurement and How Did They Find Out?

Boothroyd, R. A, McMorris, R. F., & Pruzek, R. M.[1]

State University of New York at Albany

Between 90-95% of teachers regularly construct their own tests to assess students' competency (Dorr-Bremme, 1983; Gullickson, 1982; Newman, 1981; Stiggins & Bridgeford, 1985). Despite the frequency with which teachers develop and administer tests, few states require them to complete coursework or demonstrate competency in measurement for teaching certification (Burke, 1985; Goddard, 1986). In a survey of all states and the District of Columbia, O'Sullivan and Chalnick (1991) found fewer than a third of the 51 agencies required coursework or competencies in educational measurement for initial certification. For recertification only one agency required measurement training.

Further, at least 40-60% of teacher-education programs have no course requirements in tests and measurement (Roeder, 1972, 1973; Lissitz, Schafer, & Wright, 1986; Schafer & Lissitz, 1987, 1988). Stiggins and Conklin (see Stiggins, 1991) found fewer than half the teacher training programs they surveyed even offered assessment training and fewer than a quarter of the programs required participation. Given the frequency with which teachers use self- developed tests to evaluate students, and given the paucity of requirements related to developing their measurement competency, some educators and measurement specialists question the adequacy of teachers' training in and knowledge of measurement principles.

The principal goals of this study were to examine, for a sample of science and mathematics teachers, their knowledge and skills concerning educational measurement, and to relate such knowledge and skills to selected teacher characteristics and measurement training. The approach was based on interviews and questionnaires as well as comprehensive statistical analyses using newly developed prediction methods.

Research Questions

Three questions were addressed:

1) What was the nature and extent of measurement training for these science and mathematics teachers?

2) What measurement knowledge and skills did these teachers possess?

3) What teacher characteristics are related to their measurement knowledge?

[1]We thank the 41 teachers for participating in the study, the many students in various measurement seminars for helping refine and analyze instruments as well as setting passing scores, the proposal reviewers for suggesting improvements, Angela Brayden for organizing the standards classifications, and Dr. Vicky L. Kouba for making many contributions to the dissertation from which the paper was developed.

METHOD

Sample

Seventh- and eighth-grade science and mathematics teachers were selected for the study because prior surveys indicate that classroom testing occurs with the greatest frequency within these grades and subjects.

Strong efforts were undertaken to obtain a sample that met prespecified criteria (e.g., developed their own classroom tests) yet varied in terms of the independent variables of this study (e.g., content area, experience, and type of school). Names of potential participants were obtained from a variety of sources including graduate courses at local colleges and universities, local school districts, directors of teacher centers, teachers, and friends. Teachers were screened by telephone to ensure that they were either provisionally or permanently state-certified in either 7th and 8th grade science and/or mathematics, were teaching within their certification, and had primary responsibility for constructing their own classroom tests.

The 41 participating teachers represented 25 public and private schools districts from many geographic regions in the state. No more than two teachers were selected from any one district with one exception in which four teachers were included. The districts were quite varied and included public (88%) and private (12%) schools in urban, suburban, and rural settings.

Twenty-three teachers (56%) taught 7th and 8th grade science while 18 taught mathematics at this level (44%). Approximately two-thirds (68%) were permanently state certified in their discipline while 13 (32%) had provisional certification. Female teachers outnumbered males by nearly a two-to-one margin (63% to 37%, respectively). The degree of teaching experience was somewhat evenly distributed, averaging 12 years but quite variable (SD = 7.2 years).

Instruments

Four instruments were developed for this study: (1) the Teacher Biographic Questionnaire, (2) the Measurement Competency Test, (3) the Item Judgment Task, and (4) an interview protocol.

<u>Teacher Biographic Questionnaire</u>. This 45-item questionnaire contained three sections designed to identify: (1) educational experiences appropriate to measurement, (2) classroom testing practices, and (3) attitudes toward testing.

<u>Measurement Competency Test (MCT)</u>. A 65-item, four-option, multiple-choice test was developed to assess teachers' knowledge of various measurement concepts specific to classroom testing. The test included items on test planning, types of items, item writing, reliability, and validity.

Test development began by identifying the measurement topics measurement specialists indicate are necessary for beginning teachers (e.g., Mayo, 1964; Stetz and Beck, 1978; Frisbie & Friedman, 1987; Gullickson & Hopkins, 1987). A table of specifications was constructed and a preliminary 85-item form developed using newly constructed items and adapting others from previous studies (Mayo, 1967; Newman, 1981; NCME, 1962).

A revised 80-item form was administered to 37 Masters-level students in two test construction courses. Items with extremely high or low difficulty and/or poor discrimination were revised or eliminated.

The content of the resulting 65-item test may be displayed in at least two ways. First, a content outline/table of specifications was constructed by the first author; categorizations of the items were validated, during test development, by seven doctoral-level students enrolled in a measurement seminar who classified the items into specific content domains. Additionally these students rated each item on importance of the item content for classroom teachers and quality of the item construction.

Second, items were also classified, a posteriori, by two of the authors and ten advanced graduate students according to the recent Standards for Teacher Competence in Educational Assessment of Students (AFT, NCME, & NEA, 1990). Raters assigned a 2 for an item judged relevant to a particular standard, 1 for partial relevance to a standard, and x for irrelevance to all seven standards. Due to the general nature of the standards, many items had relevance to multiple standards; however, raters were limited to assigning at most one 2 and two positive ratings per item.

The two-way classification of the MCT items (i.e., content by standard) is summarized in Table 1. As expected, the items "loaded" heavily on the first three standards, and no item was judged irrelevant by more than one rater. (Brief descriptions of the Standards maybe found in Table 5.) Apriori we had established six as the minimum sum for considering an item relevant to a standard. Each item attained at least a seven for a standard, and 45 of the 65 items had a rating sum of at least 12, as may be seen from Table 2. Approximately half of the items were judged relevant to at least two standards. The most popular standard/item intersections were for Standards 3 (Administering, scoring, and interpreting the results of both externally-produced and teacher-produced assessment methods), 2 (developing assessment methods appropriate for instructional decisions), and 1 (choosing assessment methods appropriate for instructional decisions). The three standards were ordered consistently according to the number of relevant items whether an item could be classified as fitting at most one or at most two standards.

For the 41 teachers' responses to the final 65-item test the item difficulties were somewhat evenly distributed. Twenty items (31%) were relatively easy ($p > .7$), 23 items (35%) were moderately difficult (.4 to .7), and 22 items (34%) proved difficult ($p < .4$). All but two items had positive item discriminations, with 51% (33 items) having discrimination indices above .33. (Discrimination was defined by the difference in proportions correct between the upper and lower thirds).

Item Judgment Task (IJT). Teachers reviewed 32 multiple-choice and completion items related to junior high school science and mathematics, identifying items considered "good" items and items perceived as "poor" items. Violations of recommended item writing principles (flaws) were introduced into three-quarters of the items. Items were adapted from the mathematics and science sections of the Stanford Achievement Test, the Iowa Tests of Educational Development, and the state's Junior High School Science and Mathematics Tests. The 32 items were equally divided between mathematics and science, and further faceted to include an equal number of multiple-choice and completion items. Within each of the four resulting cells, 3/4 of the items (12 of 16) contained a "flaw" in item construction.

Six types of flaws were included, three in multiple-choice items and three others in completion items. Multiple-choice flaws included: (1) a cue repeated in both stem and answer, (2) the longest, most detailed option as the keyed response, and (3) options lacking homogeneity and plausibility. Flaws incorporated in completion items included: (1) blanks in either the beginning or middle of the statement, (2) nonspecific responses as possible correct answers, and (3) omission

of a nonessential word, such as a verb.

The IJT was pilot tested with students enrolled in a graduate-level measurement course. Students reviewed each item, indicated whether it contained a flaw, and, for those items perceived as flawed, provided an explanation of the flaw. Items low in agreement between authors' intent and students' perception were revised or replaced. Illustrative items are displayed in Table 3.

Item analysis on teachers' responses to the IJT items showed that the greatest proportion of items (14 items/44%) were easy ($p > .7$), five items (16%) were moderately difficult (.4 to .7), and 41% (13 items) were difficult ($p < .4$). Two of the items had negative discrimination values and 12 items (38%) had discrimination indices less than .1. Twelve items (38%) had discrimination levels greater than .33.

Interview Protocol. An interview protocol included questions about the teacher's classroom testing practices and test development procedures [11 items], his/her measurement training [5 items], school/district policies and/or regulations specific to testing [4 items], and criteria the teacher used when identifying item flaws on the IJT [3 items].

Procedure

After prescreening, a continuous three and one-half hour block of time was scheduled with each participating teacher. The first author administered the instruments individually.

RESULTS

1. What was the nature and extent of measurement training for these science and mathematics teachers?

Approximately half (49%) of the teachers had completed at least one measurement course: 39% had taken one course while 10% completed multiple courses. The greatest number of measurement courses taken by any teacher was four. Seven teachers (12%) had completed their measurement coursework solely at the undergraduate level, 32% at the graduate level, while five percent had completed measurement courses at both levels. Comparing various groups showed that more mathematics teachers (62%) compared to science teachers (38%) reported completing a least one course in measurement. Permanently certified teachers were more likely to have completed measurement courses (57%) than were provisionally certified teachers (31%). A greater proportion of public school teachers (50%) had completed measurement training as compared to private school teachers (40%). None of these comparisons, however, was statistically significant ($p < .05$).

During the interviews, a majority of the 20 teachers who had taken measurement courses (65%) recalled that much of the content presented focused on standardized testing. For example, many teachers recollected course content dealing with derived scores (e.g., stanines and grade equivalents) and how to interpret them. Only three teachers (15%) recalled critiquing classroom test items or actually constructing tests. Most of those who had completed measurement courses estimated that only a small proportion of the course content directly assisted them in test construction.

2. What measurement knowledge and skills did these teachers possess?

Measurement Competency Test. Teachers averaged 34 items correct (53%) on the MCT out of a total of 65 items but scores varied widely (SD = 8.0), ranging from 34% correct (22 items) to 83% correct (54 items).

Subscores on the MCT were calculated for each teacher to estimate teachers' knowledge of specific measurement topics. Means ranged from 62 to 84% correct on items in content domains related to objectives, type of items, item writing, test construction, and grading and marking. Their performance was comparatively poorer (22 to 53% correct) on items specific to item analysis, standard error, and correlation, as noted in Table 4.

Subscores on the MCT were also calculated for teachers to estimate their knowledge related to specific standards (AFT, NCME, & NEA, 1990) (See Table 5). Given only one and two items "loaded" on standards 4 and 5 respectively, extreme caution should be exercised when reviewing the results presented for these two standards. Teachers correctly answered an average of 63% of items related to developing assessment methods (i.e., Standard 2) while their performance was much poorer on items relevant to choosing assessment methods (i.e., Standard 1; 42% correct) and administering, scoring, and interpreting externally-produced and teacher-produced assessment methods (i.e., Standard 3; 47% correct).

To help interpret the adequacy of teachers' performance on the MCT, 10 advanced students in a doctoral-level measurement seminar performed a modified-Angoff procedure on the MCT items. These results are also summarized in Tables 4 and 5. The average proportion for these 65 items was .54. These item proportions reflect the estimated probability of a teacher with "minimal competence in measurement" correctly answering the item. The average of these item proportions can be considered a standard denoting the minimal level of acceptable performance on these items. Overall, 44% of the teachers met or exceeded the modified-Angoff standard established by the ten judges. For MCT subscores defined according to content, the proportion of teachers meeting or exceeding the modified-Angoff subscore standards ranged from 29% to 90%.

Modified-Angoff standards for teachers' performance related to the AFT, NCME, & NEA (1990) standards are summarized in Table 5. Teachers' best performance occurred on items relevant to developing assessment methods (i.e., Standard 2) where 58% of the teachers met or exceeded the modified-Angoff standard while their worst performance was on choosing assessment methods (i.e., Standard 1) where only 34% of the teachers met or exceeded the modified-Angoff standard.

Item Judgment Task. Teachers' knowledge of measurement was also estimated using the Item Judgment Task; teachers' performance on the IJT is summarized in Table 6. On average, the teachers were able to categorize appropriately items as flawed or nonflawed and to identify correctly the type of flaw about half the time (17 items; 53%), which is the same as their average score on the MCT (53%).

IJT subscores were calculated for each teacher based on the type of item flaw, item format, and content area. Teachers seldom detected "cue in the stem" flaws in multiple-choice items (6% of judgments correct); they exhibited the greatest ability to detect the "request for a nonessential word" flaw in the completion type items (76% correct judgments). Interestingly, five teachers (12%) who detected cues in multiple-choice items identified them as a positive item characteristic,

believing cues assist less able students in obtaining correct answers. Relative difficulty levels indicate teachers more frequently detected flaws in the completion items (63% correct) as compared to flaws in multiple-choice items (43% correct). Item content, however, made little difference: 56% of the mathematics items and 51% of the science items were correctly categorized.

3. What teacher characteristics are related to their measurement knowledge?

Correlations: The correlation matrix for eleven of variables examined in this study are presented in Table 7. These variables are: 1) certification status, 2) subject taught, 3) years of teaching experience, 4) number of measurement courses completed, 5) gender, 6) frequency of classroom testing, 7) amount of time spent developing classroom tests, 8) self-report rating of measurement knowledge, 9) self-report rating of measurement training, 10) total score on the MCT, and 11) total score on the IJT.

Regression Analyses: Two stepwise multiple regression analyses were used to identify which teacher characteristics variables were useful predictors of teachers' composite scores on the MCT and the IJT. Teachers' composite scores on the MCT were regressed on 17 teacher characteristics. The number of measurement courses entered in the first step, accounting for 26% of the variance in teachers' MCT scores. In step two, teachers' self-report rating of adequacy of measurement training accounted for an additional 8% of the variance. In step three, teachers' rating of their level of measurement knowledge accounted for an additional 8% of the variance. Overall, the three predictors accounted for 42% of the variance in teachers' scores on the MCT.

Teachers' self-report ratings of their measurement knowledge may have acted as a suppressor variable; these ratings had virtually no correlation (-.01) with teachers' MCT scores, but had moderate correlations with the other two predictors, number of measurement courses (.31) and teachers' perception regarding adequacy of their measurement training (.40).

In the second regression, teachers' IJT scores were regressed on the same 17 predictors used in the previous analysis with one additional predictor: teachers' MCT scores. Two predictors, teachers' MCT scores and self-reported measurement competency, explained 42% of the variance in teachers' IJT scores.

Interbattery Factor Analysis: A newly developed form of canonical, interbattery, and regression analysis (Pruzek, 1992) was employed to study the relationships between a set of predictor and criterion variables. This new method involves computing interbattery factor coefficients to account for the 'cross-battery' correlations and errors of measurement. The 'cross battery' correlations are then linked to the canonical structure matrices vis-a-vis canonical variate analysis (cf. Browne, 1979). However, the entire process of estimation is begun from a joint convex sum covariance (correlation) matrix, following the logic and general procedures described in detail by Pruzek and Lepak (1991). The resulting interbattery factor coefficients are presented in Table 8 and are used as a basis for accounting for relationships between the two sets of variables. The two sets of variables were defined as follows: Set one, considered as predictor measures, consisted of nine variables which were: 1) certification status, 2) subject taught, 3) years of teaching experience, 4) number of measurement courses completed, 5) gender, 6) frequency of classroom testing, 7) amount of time spent developing classroom tests, 8) self-report rating of measurement knowledge, and 9) self-report rating of measurement training. Set two, considered as criterion measures, consisted of 20 variables representing teachers' scores on the 13 subscores of the MCT and the 7 subscores of the IJT (see Table 8).

The five-factor solution accounted for 42% of the total variance. Interpretation of these factors was based on examining variables with loadings greater than |.35|. Using this criterion, the five factors can be summarized as follows.

Factor I illustrates the relationships among teachers' knowledge on five subsets of MCT items (types of tests, test construction, item analysis, correlation, and standard error), their ability to detect nonhomogeneity of distractors and misplaced blanks in IJT items with completion of more measurement courses, feeling more adequately trained in measurement, and being male.

The relationships in Factor II are among performance on three subsets of MCT items (reliability, standard error, and validity), and an inability to detect the "Cues in the stem" flaw in the IJT items, with less teaching experience, more frequently testing, spending more time developing tests, and being female.

The third factor indicates knowledge of item analysis relates with permanent certification status, teaching mathematics, teaching experience, and feeling less adequate in measurement knowledge.

Factor IV indicates knowledge in five of the MCT topics (test planning, objectives, types of tests, item analysis, and score interpretation), and ability to detect the longest option and nonspecific response flaws in IJT items are related with feeling more adequately trained in measurement and being female.

The relationship demonstrated in the fifth factor links knowledge on three MCT subsets (score interpretation, grading & marking, validity) with being male.

DISCUSSION

Consistent with findings from previous studies, results from this study indicate that 7th and 8th grade mathematics and science teachers frequently test students with teacher-made tests approximately one test every two weeks per class. Teachers were found to place more weight on students' scores on these tests when assigning end-of-course grades than on other forms of assessment. The frequency with which teachers administer self-developed tests and their heavy reliance on these tests in assigning course grades raises questions regarding the extent to which teachers have the measurement knowledge and skills necessary to construct and interpret effective classroom tests.

Based on these results it can be inferred that teachers' knowledge of measurement is not sufficient. For example, nearly 56% of the teachers scored below a modified-Angoff standard on the MCT. Teachers' deficiencies in measurement knowledge probably result at least partially from inadequate training given that 51% of the teachers never completed a single measurement course. Although additional unspecified measurement coursework would probably enhance teachers' ability to construct effective classroom tests, courses specifically devoted to teacher-made tests and teachers' use of information for instruction and for grading may be especially warranted. This conclusion is supported by findings from the regression analyses. For example, teacher competence shown on the MCT was predicted by the number of measurement courses. Similarly, Plake, Impara, and Fager (1992) found that teachers who completed a course or inservice training program in measurement had higher scores on a competency test than did those without such background. Such data are supportive of the value of measurement training although admittedly

not ironclad proof of causality.

The infrequency of appropriate coursework in measurement required of or taken by teachers, and the less-than-stellar measurement competencies displayed by teachers in this study and elsewhere, contrast with the needs for teachers to develop and interpret information appropriate to instruction. Merwin (1989) indicates that teachers make numerous decisions about students and programs on a daily basis, the quality of those decisions are dependent on teachers' abilities to effectively identify and evaluate important characteristics. Some examples from Reynolds (1992) list of instructional tasks that require teachers' decisions include:

> Implementing and adjusting plans during instruction;
> Organizing and monitoring students, time, and materials during instruction;
> Evaluating student learning; and
> Reflecting on one's own actions and students' responses in order to improve teaching. (p. 4)

Such tasks are required for competent instruction, and yet teachers are not routinely instructed in ways to collect and interpret information. Crucial ways to collect information certainly include classroom testing. Most teachers have not been adequately trained in how to develop and interpret a classroom test, even though these tests are the primary basis for assigning course grades and a major basis for a plethora of educational outcomes.

Many additional ways to collect and use information for instructional purposes are specified by Reynolds (1992), Airasian (1991), and others. Airasian (1991) underlines teachers' use of information to size up new pupils, plan instruction, critique instructional materials, estimate how instruction is going, and so on. Much of the information is collected informally by teachers who have had little guidance from the measurement community in how to consider the validity and reliability of such information.

The responsibility for ensuring that teachers have the requisite knowledge and skills to perform these tasks effectively must be shared. Merwin (1989) suggests that teacher training programs have a professional obligation to ensure teachers possess adequate assessment skills. The AFT, NCME, and NEA (1990) standards on teacher competency in measurement reflect a very positive and significant step toward this goal. Airasian (1991) provides a challenge for instructors in measurement to develop courses that include expanding on and illustrating the measurement concepts we hold dear in situations even more informal than the teacher-developed paper-and-pencil test. We believe that only through a cooperative undertaking by education and measurement communities can a significant change be made in teachers' classroom assessment practices. Although the AFT, NCME, and NEA (1990) standards provide a broad framework within which change should occur, specific steps need to be identified that will provide guidance on how these standards can be transformed to realizations.

REFERENCES

Airasian, P. W. (1991). Perspectives on measurement instruction. Educational Measurement: Issues and Practice, 10(1), 13-16,26.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association (1990). Standards for teacher competence in educational assessment of students. Washington, DC: NCME.

Boothroyd, R. A. (1990). Variables related to the characteristics and quality of classroom tests: An exploratory study with seventh and eighth grade science and mathematics teachers. (Doctoral Dissertation, The University at Albany, 1990) Dissertation Abstracts International, 51/07A, 2355.

Browne, M. W. (1979). The maximum likelihood solution in interbattery factory analysis. British Journal of Mathematical and Statistical Psychology, 32, 75-86.

Burke, M. P. (1985). Requirements for certification. (5th ed.). Chicago, IL: The University of Chicago Press.

Dorr-Bremme, D. W. (1983). Assessing students: Teachers' routine practices and reasoning. Evaluation Comment, 6, 1-12.

Frisbie, D. A., & Friedman, S. J. (1987). Test standards -- Some implications for the measurement curriculum. Educational Measurement: Issues and Practice, 6(3), 17-23.

Goddard, R. E. (1986). Teacher certification requirements. (4th ed.). Sarasota, FL: Teacher Certification Publication.

Gullickson, A. R. (1982). The practice of testing in elementary and secondary schools. Paper presented at the Rural Education Conference at Kansas State University, Manhattan, KA. (ERIC Document Reproduction Service No. ED 229 391).

Gullickson, A. R., & Hopkins, K. D. (1987). The context of educational measurement instruction for preservice teachers: Professional perspectives. Educational Measurement: Issues and Practice, 6(3), 12-16.

Lissitz, R. W., Schafer, W. D., & Wright, M. V. (1986, April). Measurement training for school personnel: Recommendations and reality. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

Mayo, S. T. (1964). What experts think teachers ought to know about educational measurement. Journal of Educational Measurement, 1, 79-86.

Mayo, S. T. (1967). Preservice preparation of teachers in educational measurement. Final Report Project No. 5-0807, Contract No. OE4-10-011. Washington, DC: United States Office of Education, and Loyola University.

Merwin, J. C. (1989). Evaluation. In M. C. Reynolds (Ed.), Knowledge base for the beginning teacher. (pp. 185-192). Oxford: Pergamon Press.

National Council on Measurement in Education. (1962). Multiple-choice items for a test of teacher competence in educational measurement. Washington, DC: Author.

Newman, D. C. (1981). Teacher competency in classroom testing, measurement preparation, and classroom testing practices. (Doctoral dissertation, Georgia State University). Dissertation Abstracts International, 45(2), 1111A.

O'Sullivan, R. E., & Chalnick, M. K. (1991). Measurement-related coursework requirements for teacher certification and recertification. Educational Measurement: Issues and Practice, 10,(1) 17-19,23.

Plake, B. S., Impara, J. C., & Fager, J. J. (1992, April). Assessment competencies of teachers: A national survey. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

Pruzek, R. M., & Lepak, G. M. (1991). Weighted structural regression: A broad class of adaptive methods for improving linear prediction. Multivariate Behavioral Research, 27(1), 95-129.

Pruzek, R. M. (1992). Personal Communication.

Reynolds, A. (1992). What is competent beginning teaching? A review of the literature. Review of Educational Research, 62, 1-35.

Roeder, H. H. (1972). Are today's teachers prepared to use tests? Peabody Journal of Education, 49, 239-240.

Roeder, H. H. (1973). Teacher education curricula--your final grade is F. Journal of Educational Measurement, 10, 141-143.

Schafer, W. D., & Lissitz, R. W. (1987). Measurement training for school personnel: Recommendations and reality. Journal of Teacher Education, 38(3), 57-63.

Schafer, W. D., & Lissitz, R. W. (1988, March). The current status of teacher training in measurement. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Stetz, F. P., & Beck, M. D. (1978, April). A survey of opinions concerning users of educational tests. Paper presented at the meeting of the National Council on Measurement in Education, Toronto, Ontario

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. Journal of Educational Measurement, 20, 271-286.

Stiggins, R. J. (1991). Relevant classroom assessment training for teachers. Educational Measurement: Issues and Practice, 10,(1) 7-12.

Table 1

Classification of MCT Items by Content Domain and Standard[a]

Content Domain	Standard Number[b]							Total Number of Items
	1	2	3	4	5	6	7	
1. Test Preparation		2						2
2. Objectives		7	1					8
3. Types of Tests[c]	2	1	4	1				7
4. Types of Items	4	3						7
5. Writing Items		4						4
6. Test Construction		3						3
7. Item Analysis		2	5					7
8. Score Interpretation			6					6
9. Grading & Marking			3		2			5
10. Correlation			3					3
11. Reliability			6					6
12. Standard Error			2					2
13. Validity	1	1	3					5
Total Number of Items[c]	7	23	33	1	2			65

[a]The standards are from AFT, NCME, & NEA, 1990.

[b] The rating sums are based on 12 raters each assigning a 2 for an item judged relevant to a standard, a 1 for partial relevance, and an x for irrelevance to all Standards. (Each rater could assign at most one 2 and two positive ratings per item.) Items were assigned to the standard with the highest rating sums.

[c]One item had the identical sum rating on standards 3 and 4.

Table 2

MCT Items Related to Standards[a]

Rating Sums	Standard 1	2	3	4	5	6	7	Number of Sums
21-23			1					1
18-20		3	6		1			10
15-17	3	5	8					16
12-14	4	5	8	1				18
9-11	3	10	9	2				24
6-8	9	9	6	4	2			30
Number of Items	19	32	38	7	3			99

Note. The rating sums are based on 12 raters each assigning a 2 for an item judged relevant to a standard, a 1 for partial relevance, and an x for irrelevance to all Standards. (Each rater could assign at most one 2 and two positive ratings per item.)

Intersections with sums greater than six are included in this table; approximately half of the items were related to at least two standards. For items judged as related to multiple standards, only the two highest intersections were identified.

[a]The standards are from AFT, NCME, & NEA, 1990.

Table 3

Sample Flawed Items from the Item Judgment Task

Multiple-choice science item; longest and more detailed response as the key

Sunburn is caused by
a. heat.
*b. the ultraviolet rays present in daylight.
c. visible light.
d. wind.

Multiple-choice mathematics item; non-homogeneity of distractors

A television set was originally priced at $204. It is now on sale for $153. The price of the television has been reduced
*a. twenty-five percent.
b. 1/5.
c. $55.00.
d. 75%.

Completion mathematics item; Non-specific answer

A square is a specific form of (quadrilateral).

Completion science item; Request for a non-essential word

As a storm approaches, decreased air pressure usually causes a drop in a barometer's (mercury).

Note. * and (____) represent the keyed or desired response.

Table 4

Teachers' Performance on the MCT by Content Domain

Content Domain	Number of Items	Percentage of Items Correct	Average Number of Items Correct	Standard Deviation	Modified Angoff Standard	Percentage of Teachers *At or Above* the Standard
1. Test Preparation	2	44	.88	.71	.80	68
2. Objectives	8	62	5.00	1.67	4.66	58
3. Types of Tests	7	40	2.83	1.48	3.08	34
4. Types of Items	7	56	3.93	1.37	4.29	32
5. Writing Items	4	85	3.39	.67	3.09	90
6. Test Construction	3	63	1.87	.69	1.87	76
7. Item Analysis	7	38	2.66	1.49	3.03	29
8. Score Interpretation	6	54	3.27	1.45	3.36	44
9. Grading & Marking	5	62	3.12	1.12	3.27	27
10. Correlation	3	53	1.59	.81	1.76	49
11. Reliability	6	46	2.76	1.16	2.88	58
12. Standard Error	2	22	.44	.63	.55	37
13. Validity	5	50	2.49	.78	2.57	42
Total Test	65	53	34.23	8.08	35.21	44

Table 5

Teachers' Performance on the MCT by Standard[a]

Standard	Number of Items	Percentage of Items Correct	Average Number of Items Correct	Standard Deviation	Modified Angoff Standard	Percentage of Teachers *At or Above* the Standard
1. choosing assessment methods	7	42	2.97	1.49	3.38	34
2. developing assessment methods	23	63	14.60	3.27	14.08	58
3. administering, scoring, and interpretation	33	47	15.37	4.71	16.44	42
4. using assessment results in making decisions	1	46	.46	.51	.50	46
5. developing a valid grading system	2	58	1.17	.74	1.27	56
6. communicating assessment results	0	N/A	N/A	N/A	N/A	N/A
7. recognizing inappropriate use of assessment results	0	N/A	N/A	N/A	N/A	N/A
Total Test[b]	65	53	34.23	8.08	35.67	43.9

[a]The standards are from AFT, NCME, & NEA, 1990.

[b]One item had the identical sum rating on standards 3 and 4 and was included on both.

Table 6

Teachers' Performance on the Item Judgment Task by Item Flaw, Format, and Content

Type of Flaw	Number of Items	Percentage of Items Correct	Average Number of Items Correct	Standard Deviation
1. Nonflawed (MC & C)	8	83	6.66	1.15
2. Cue in Stem (MC)	4	6	.27	.45
3. Nonhomogeneity of Distractors (MC)	4	46	1.85	1.04
4. Keyed Response the Longest Option (MC)	4	34	1.34	1.51
5. Nonspecific Response (C)	4	54	2.17	1.38
6. Inappropriately Placed Blank (C)	4	42	1.66	1.11
7. Nonessential Word (C)	4	76	3.02	.69
Type of Item				
1. Multiple-choice	16	43	6.88	2.28
2. Completion	16	63	10.10	2.05
Item Content				
1. Mathematics	16	56	8.90	2.13
2. Science	16	51	8.07	1.72
Total Task	32	53	16.98	3.23

<u>Note</u>. MC denotes flaws in the multiple-choice items while C denotes flaws in the completion items.

Table 7

Correlation Matrix

Variable	CS	ST	TE	MC	G	FoT	TDT	MK	AT	MCT
Certification status (CS)										
Subject taught (ST)	.29									
Teaching experience (TE)	.49	-.04								
Measurement coursework (MC)	.21	-.06	.24							
Gender (G)	.19	-.06	.50	.18						
Frequency of testing (FoT)	-.06	.19	.10	-.34	-.14					
Test development time (TDT)	-.27	.14	-.27	.24	.27	-.06				
Measurement knowledge (MK)	-.33	.18	-.25	-.31	.08	.16	.10			
Adequate training (AT)	.11	.30	.19	.40	-.18	-.01	.16	-.29		
MCT Score (MCT)	.12	.18	-.02	.50	.06	-.16	.21	.01	.49	
IJT Score (IJT)	.21	.09	.09	.56	-.14	-.07	.01	-.38	.35	.53

These variables were coded as follows

Certification status: 0=Provisional; 1=Permanent
Subject taught: 0=Mathematics; 1=Science
Gender: 0=Male; 1=Female
Frequency of testing: 1=Several times a week; 2=Once a week; 3=Several times a month; 4=Once a month; 5=Several times a year; 6=Once a year; 7=Never
Measurement knowledge: 1=Excellent; 2=Very Good; 3=Good; 4=Adequate; 5=Poor
Adequate training: 1=Strongly Disagree; 2=Disagree; 3=Uncertain; 4=Agree; 5=Strongly Agree

Table 8

Rotated Interbattery Factor Coefficients[a] Based on Convex Sums

Variable / Factor	I	II	III	IV	V	h^2
Certification Status[b]	-.01	-.18	**.74**	.11	.22	.65
Subject Taught[b]	.09	-.31	**-.36**	.42	-.08	.41
Teaching Experience	.23	**-.50**	**.52**	-.12	.21	.64
T & M Courses	**.75**	.17	.17	.13	.22	.68
Gender[b]	**-.40**	**.65**	-.10	**.36**	**-.56**	1.00
Testing Frequency[c]	-.18	**-.53**	-.15	.13	.07	.36
Time Developing Tests	.15	**.46**	-.31	.06	.10	.34
Measurement Knowledge[c]	-.34	-.09	**-.52**	.04	.12	.40
Adequacy of Training	**.39**	.02	.12	**.35**	.25	.35
Test Planning (MCT)	.15	-.08	-.14	**.48**	.17	.31
Objectives	.01	.04	.05	**.62**	-.02	.39
Types of Tests	**.49**	.25	-.07	.21	.18	.38
Types of Items	.15	.12	-.02	**.47**	.07	.26
Item Writing	.20	-.02	.29	.12	-.05	.14
Test Construction	**.62**	.00	-.06	.25	-.08	.45
Item Analysis	**.41**	.09	**.47**	**.58**	.08	.73
Score Interpretation	.00	.21	-.03	**.54**	**.44**	.53
Grading & Marking	.30	-.18	.05	.31	**.52**	.49
Correlation	**.42**	.31	-.32	.15	.12	.41
Reliability	.08	**.39**	-.10	.05	.12	.18
Standard Error	**.54**	**.42**	-.08	.18	.07	.51
Validity	.03	**.51**	.04	.25	**.53**	.60
Nonflawed Items (IJT)	-.05	.02	.27	-.22	.01	.12
Cue in Stem	.19	**-.54**	-.07	.28	-.10	.42
Nonhomogeneity	**.50**	-.05	.00	.10	-.10	.28
Longest Option	.34	-.01	-.07	**.48**	.01	.35
Nonspecific Response	.28	-.02	.14	**.37**	.13	.26
Misplaced Blank	**.43**	.27	.05	.31	.14	.38
Nonessential Word	.12	.24	.20	-.17	.10	.15
Proportion of Variance	.11	.09	.07	.10	.05	.42

[a]Loadings greater in magnitude than |.35| are in **boldface**.

[b]These binary variables are coded as follows
Certification Status: 0=Provisional; 1=Permanent
Subject Taught: 0=Mathematics; 1=Science
Gender: 0=Male; 1=Female

[c]The response scales on these items are such that a negative loading reflects higher levels of testing and perceived measurement knowledge.